Designing an Efficient Multimodal Biometric System using Palmprint and Speech Signal

Mahesh P.K.[1], M.N. Shanmukha Swamy[2]
[1] JSS Research Foundation, SJCE, Mysore, India
Email: mahesh24pk@gmail.com
[2] JSS Research Foundation, SJCE, Mysore, India
Email: mnsjce@gmail.com

***Abstract*— This paper proposes a multimodal biometric system using palmprint and speech signal. In this paper, we propose a novel approaches for both the modalities. We extract the features using Subband Cepstral Coefficients for speech signal and Modified Canonical method for palmprint. The individual feature score are passed to the fusion level. Also we have proposed a new fusion method called weighted score. This system is tested on clean and degraded database collected by the authors for more than 300 subjects. The results show significant improvement in the recognition rate.**

***Index Terms*—Multimodal biometrics, Speech signal, Palmprint, Fusion**

I. Introduction

A unimodal biometric authentication, which identifies an individual person using physiological and/or behavioral characteristics, such as palmprint, face, fingerprints, hand geometry, iris, retina, vein and speech. These methods are more reliable and capable than knowledge-based (e.g. Password) or token-based (e.g. Key) techniques. Since biometric features are hardly stolen or forgotten.

However, a single biometric feature sometimes fails to be exact enough for verifying the identity of a person. By combining multiple modalities enhanced performance reliability could be achieved. Due to its promising applications as well as the theoretical challenges, multimodal biometric has drawn more and more attention in recent years [1]. Speech Signal and palmprint multimodal biometrics are advantageous due to the use of non-invasive and low-cost speech and image acquisition. In this method we can easily acquire palmprint images using digital cameras, touchless sensors and speech signal using microphone. Existing studies in this approach [2, 3] employ holistic features for palmprint and speech signal representation and results are shown with different techniques of fusion and algorithms.

Multimodal system also provides anti-spooling measures by making it difficult for an intruder to spool multiple biometric traits simultaneously. However, an integration scheme is required to fuse the information presented by the individual modalities.

This paper presents a novel fusion strategy for personal identification using speech signal and palmprint features at the features level fusion Scheme. The proposed paper shows that integration of speech signal and palmprint biometrics can achieve higher performance that may not be possible using a single biometric indicator alone. We extract the features using modified canonical form method for palmprint and Subband Cepstral Coefficients for speech. Integrating these two features at fusion level, which gives better performance and better accuracy. Which gives better performance and better accuracy for both traits (speech signal & palmprint).

The rest of this paper is organized as fallows. Section 2 presents the system structure, which is used to increase the performance of individual biometric trait; multiple classifiers are combined using matching scores. Section 3 presents feature extraction method used for speech signal and section 4 for palmprint. Section 5, the individual traits are fused at matching score level based on weighted sum of score technique. Finally, the experimental results are given in section 6. Conclusions are given in the last section.

II. System Overview

The block diagram of a multimodal biometric system using two (palm and speech) modalities for human recognition system is shown in Figure 1. It consists of three main blocks, that of Preprocessing, Feature extraction and Fusion. Preprocessing and feature extraction are performed in parallel for the two modalities. The preprocessing of the audio signal under noisy conditions includes signal enhancement, tracking environment and channel noise, feature estimation and smoothing [4]. The preprocessing of the palmprint typically consists of the challenging problems of detecting and tracking of the palm and the important palm features.

Further, features are extracted from the training and testing images and speech signal respectively, and then matched to find the similarity between two feature sets. The matching scores generated from the individual recognizers are passed to the decision module where a person is declared as genuine or an imposter.

III. Subband Based Cepstral Coefficients And Gaussian Mixture Model

A. Subband Decomposition via Wavelet Packets

A detailed discussion of wavelet analysis is beyond the scope of this paper and we therefore refer interested readers to a more complete discussion presented in [5]. In continuous time, the Wavelet Transform is defined as the inner product of a signal x(t) with a collection of wavelet functions $y_{ab}(t)$ in which the wavelet functions are scaled(by a) and translated

DOI: 01.IJSIP.03.01.7

(by b) versions of the prototype wavelet y(t).

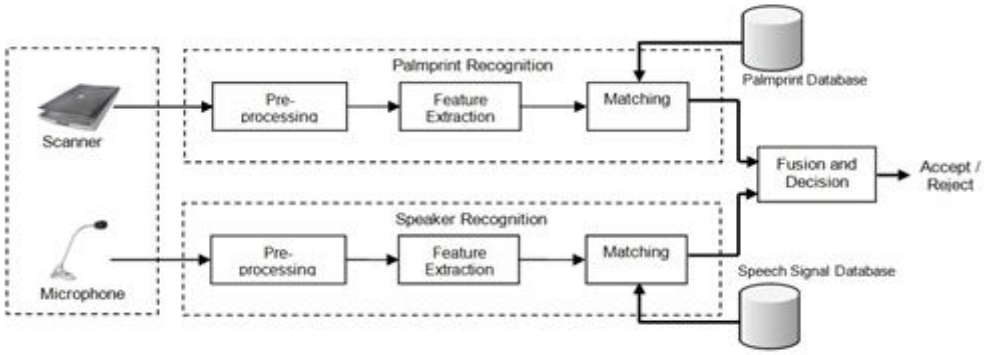

Figure 1. Block diagram of the proposed multimodal biometric verification system

$$\psi_{a,b}(t) = \psi\left(\frac{t-b}{a}\right)dt \quad (1)$$

$$W_\psi x(a,b) = \frac{1}{\sqrt{a}}\int_{-\infty}^{+\infty} x(t)\psi^*\left(\frac{t-b}{a}\right)dt \quad (2)$$

Discrete time implementation of wavelets and wavelet packets are based on the iteration of two channel filter banks which are subject to certain constraints, such as low pass and/or high pass branches on each level followed by a sub sampling-by-two unit. Unlike the wavelet transform which is obtained by iterating on the low pass branch, the filterbank tree can be iterated on either branch at any level, resulting in a tree structured filterbank which we call a wavelet packet filterbank tree. The resultant transform creates a division of the frequency domain that represents the signal optimally with respect to the applied metric while allowing perfect reconstruction of the original signal. Because of the nature of the analysis in the frequency domain it is also called subband decomposition where subbands are determined by a wavelet packet filterbank tree.

B. Wavelet Packet Transform Based Feature Extraction Procedure

Here, speech is assumed to be sampled at 8 kHz. A frame size of 24msec with a 10msec skip rate is used to derive the Subband based Cepstral Coefficients features, whereas a 20msec frame with the same skip rate is used to derive the MFCCs. We have used the same configuration proposed in [6] for MFCC. Next, the speech frame is Hamming windowed and pre-emphasized.

The proposed tree assigns more subbands between low to mid frequencies while keeping roughly a log-like distribution of the subbands across frequency. The wavelet packet transform is computed for the given wavelet tree, which results in a sequence of subband signals or equivalently the wavelet packet transform coefficients, at the leaves of the Tree. In effect, each of these subband signals contains only restricted frequency information due to inherent bandpass filtering. The wavelet packet tree is given in Figure 2. The energy of the sub-signals for each subband is computed and then scaled by the number of transform coefficients in that subband.

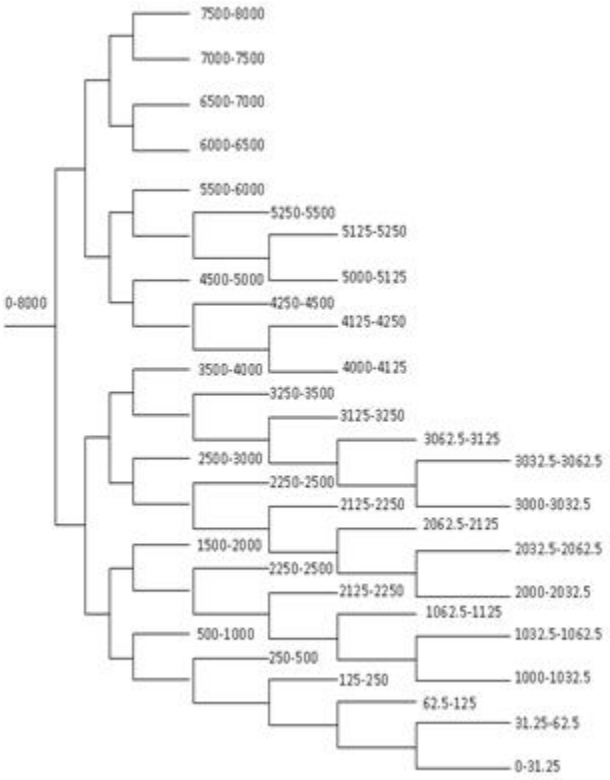

Figure 2. Wavelet Packet Tree

The subband signal energies are computed for each frame as,

$$S_i = \frac{\sum_{mel}\left[(W_\psi)(i), m\right]}{N_i} \quad (3)$$

W_ψ : Wavelet packet transform of signal x,
i :subband frequency index (i=1,2...L),
N_i : number of coefficients in the i^{th} subband.

C. Subband based Cepstral Coefficients

As in MFCCs the derivation of coefficients is performed in two stages. The first stage is the computation filterbank energies and the second stage would be the decorrelation of the log filterbank energies with a DCT to obtain the MFCC. The derivation of the Subband Based Cepstral coefficients follows the same process except that the filterbank energies

DOI: 01.IJSIP.03.01.7

are derived using the wavelet packet transform rather than the short-time Fourier transform. It will be shown that these features outperform MFCCs. We attribute this to the computation of subband signals with smooth filters. The effect of filtering as a result of tracing through the low-pass/high-pass branches of the wavelet packet tree, is much smoother due to the balance in time-frequency representation. We believe that this will contribute to improved speech/speaker characterization over MFCC. Subband Based Cepstral coefficients are derived from subband energies by applying the Discrete Cosine Transformation:

$$SBC(n) = \sum_{i=1}^{L} \log S_i \cos\left(\frac{n(i-0.5)}{L}\pi\right), n = 1,...n' \quad (4)$$

where n' is the number of SBC coefficients and L is the total number of frequency bands. Because of the similarity to root-cepstral [7] analysis, they are termed as subband based cepstral coefficients.

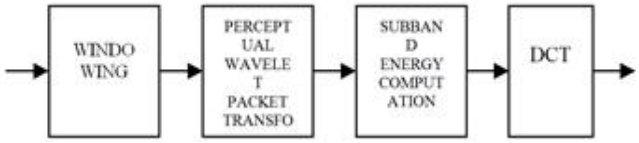

Figure 3. Block diagram for Wavelet Packet Transform based feature extraction procedure

D. The Gaussian Mixture Model

In this study, a Gaussian Mixture Model approach proposed in [8] is used where speakers are modeled as a mixture of Gaussian densities. The use of this model is motivated by the interpretation that the Gaussian components represent some general speaker-dependent spectral shapes and the capability of Gaussian mixtures to model arbitrary densities.

The Gausssian Mixture Model is a linear combination of M Gaussian mixture densities, and given by the equation,

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \quad (5)$$

Where $\vec{x}$ is a D-dimensional random vector, $b_i(\vec{x})$, $i=1,...M$ are the component densities and p_i, $i=1,\ldots M$ are the mixture weights. Each component density is a D-dimensional Gaussian function of the form

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2} \mid \sum_i \mid^{1/z}} \exp\left\{-\frac{1}{2}(\vec{x}-\vec{\mu})^T \sum_i^{-1} (\vec{x}-\vec{\mu})\right\} \quad (6)$$

Where $\vec{\mu}$ denotes the mean vector and $\sum_i$ denotes the covariance matrix. The mixture weights satisfy the law of total probability, $\sum_{i=1}^{M} p_i = 1$. The major advantage of this representation of speaker models is the mathematical tractibility where the complete Gaussian mixture density is represented by only the mean vectors, covariance matrices and mixture weights from all component densities.

IV. Feature Extraction Using Modified Canonical Form Method

Features are the attributes or values extracted to get the unique characteristics from the image and speech signal.

A. Palmprint feature extraction methodology

Details of the algorithm are as follows:

1) Identify hand image from background

Our designed system is such that palmprint images are captured using contact-less without pegs, keeping the image background relatively uniform and relatively low intensity when compared to the hand image. Using the statistical information of the background, the algorithm estimates an adaptive threshold to segment the image of the hand from the background. Pixels with intensity above the threshold are considered to be part of the hand image.

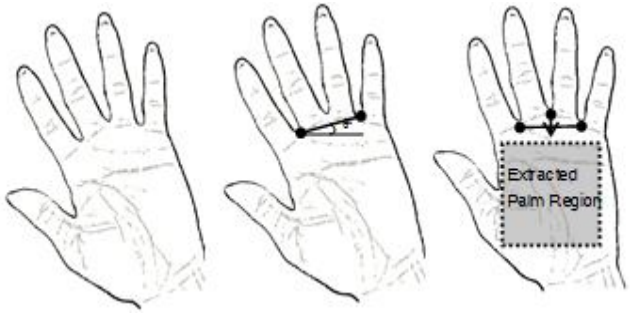

Figure 4. Schematic diagram of image alignment

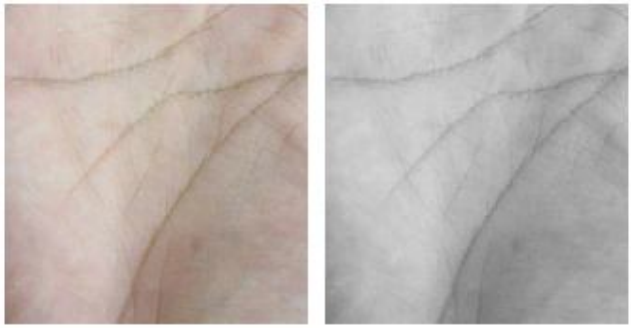

Figure 5. Segmentation of ROI

2)Locate region-of-interest

The palm area is extracted from the binary image of the hand. After translating the original image into binary image, we find two key positioning points in the palmprint image using automatic detecting method. The first valley in the graph is the gaps between little finger and ring finger, Key Point 1. The third valley in the graph is the gaps between middle finger and index finger, Key Point 2. The key point is circled in Figure 4. The hand image is rotated by θ degrees. The hand images are rotated to align the hand images into a predefined direction. θ is calculated using the key points as shown in the Figure 4. Since the size of the original image is large, a smaller hand image is cropped out from the original hand image after image alignment using key points. Figure 5

DOI: 01.IJSIP.03.01.7

shows the proposed image alignment and ROI selection method.

B. Modified Canonical Form Method

The "Eigenpalm" method proposed by Turk and Pentland [9][10] is based on Karhunen-Loeve Expression and we are motivated by this work for efficiently representing picture of images. The Eigen method presented by Turk and Pentland finds the principal components (Karhunen-Loeve Expression) of the image distribution or the eigenvectors of the covariance matrix of the set of images. These eigenvectors can be thought as set of features, which together characterized between images

Let a image *I (x, y)* be a two dimensional array of intensity values or a vector of dimension *n*. Let the training set of images be $I_1, I_2, I_3, \ldots\ldots In$. The average image of the set is defined by

$$\Psi = \frac{1}{N}\sum_{i=1}^{n} I\, i \tag{7}$$

Each image differed from the average by the vector.

$$\phi_I = I\, i - \Psi \tag{8}$$

This set of very large vectors is subjected to principal component analysis which seeks a set of *K* orthonormal vectors *Vk, K=1,……, K* and their associated eigenvalues λk which best describe the distribution of data. The vectors *Vk* and scalars λk are the eigenvectors and eigenvalues of the covariance matrix:

$$C = \frac{1}{N}\sum_{i=1}^{N} \phi_i\, \phi_i^T = A\, A^T \tag{9}$$

Where the matrix $A = [\phi_1, \phi_2 \ldots\ldots\ldots \phi_N]$ finding the eigenvectors of matrix *Cnxn* is computationally intensive. However, the eigenvectors of *C* can determine by first finding the eigenvectors of much smaller matrix of size *NxN* and taking a linear combination of the resulting vectors [4].

The modified canonical method proposed in this paper is based on Eigen values and Eigen vectors. These Eigen valves can be thought a set of features which together characterized between images.

Let $\hat{P}$ be the normalized modal matrix of I, the diagonal matrix is given by

$$D = \hat{P}^{-1}\, C\, \hat{P} \tag{10}$$

Where $\hat{P} = \frac{Vk_{ij}}{X_i}$ and

$$X_i = sqrt(\sum Vk_{ij}{}^2)\,,\ i, j=1,2,3,\ldots n \tag{11}$$

Then the quadratic form Q is given by

$$Q = X^T D\, X \tag{12}$$

The following steps are considered for the feature extraction:

- Select the palm image for the input
- Pre-process the image
- Determine the eigen values and eigen vectors of the image
- Use the canonical form for the feature extraction.

C. Euclidean Distance

Let an arbitrary instance X be described by the feature vector

$$X = [a_1(x), a_2(x) \ldots\ldots\ldots a_n(x)] \tag{13}$$

Where $a_r(x)$ denotes the value of the r^{th} attribute of instance x. Then the distance between two instances x_i and x_j is defined to be $d(x_i, x_j)$;

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} (a_r(X_i) - a_r(X_j))^2} \tag{14}$$

D. Score Normalization

This step brings both matching scores between 0 and 1 [11]. The normalization of both the scores are done by

$$N_{Speech} = \frac{MS_{Speech} - \min_{Speech}}{\max_{Speech} - \min_{Speech}} \tag{15}$$

$$N_{Palm} = \frac{MS_{Palm} - \min_{Palm}}{\max_{Palm} - \min_{Palm}} \tag{16}$$

Where $\min_{Speech}$ and $\max_{Speech}$ are the minimum and maximum scores for speech signal recognition and $\min_{Palmprint}$ and $\max_{Palmprint}$ are the corresponding values obtained from palmprint trait.

E. Generation of Similarity Scores

Note that the normalized score of palmprint which is obtained through Haar Wavelet gives the information of dissimilarity between the feature vectors of two given images while the normalized score from speech signal gives a similarity measure. So to fuse both the score, there is a need to make both the scores as either similarity or dissimilarity measure. In this paper, the normalized score of palmprint is converted to similarity measure by

$$N'_{Palm} = 1 - N_{Palm} \tag{17}$$

V. Fusion

The biometrics systems are integrated at multi-modality level to improve the performance of the verification system. At multi-modality level, matching score are combined to give a final score. The following steps are performed for fusion:

1. Given a query image and speech signal as input, features are extracted by the individual recognition and then the matching score of each individual trait is calculated.
2. The weights a and b are calculated using FAR and FRR.
3. Finally, the final score after combining the matching score

DOI: 01.IJSIP.03.01.7

of each trait is calculated by weighted sum of score technique.

$$MS_{fusion} = \frac{a * MS_{Palm} + b * MS_{Speech}}{2} \quad (18)$$

Where *a* and *b* are the weights assigned to both the traits. The final matching score (MS_{fusion}) is compared against a certain threshold value to recognize the person as genuine or an imposter.

VI. Experimental Results

This section shows the experimental results of our approach with Modified Canonical method and Subband based Cepstral coefficients for palmprint and Speech respectively. We evaluate the proposed multimodal system on a data set including more than 300 subjects taking 6 different samples, also we have experimented with two different conditions (Cleaned and Degraded data). The training database contains a palmprint images and speech signal for each individual for each subject.

The comparison of both unimodal systems (palm and speech modality) and a bimodal system is given in Table 1 & 2. It can be seen that the fusion of palmprint and speech features improves the verification score. The experiments show that EER is reduced to 3.54% in clean database and 9.17% in degraded database.

Table I. The FAR And FRR Of Palmprint And Speech Signal In Clean And Degraded Conditions

Modality	Method	Database	Classifier	FAR%	FRR%
Speech Signal	SBC	Clean database of 300 subjects	GMM	03.54	10.73
Palmprint	MCF		----	06.73	15.24
Speech Signal	SBC	Degraded database of 50 subjects	GMM	46.67	31.67
Palmprint	MCF		----	23.33	26.67

Table II. The FAR And FRR After Fusion

Methods		Database	FAR%	FRR%
Palmprint	Speech Signal			
MCF	SBC	Clean database of 300 subjects	02.76	04.32
MCF	SBC	Degraded database of 50 subjects	06.67	11.67

Conclusions

Biometric systems are widely used to overcome the traditional methods of authentication. But the unimodal biometric system fails in case of biometric data for particular trait. This paper proposes a new method in selecting and dividing the ROI for analysis of palmprint. The new method utilizes the maximum palm region of a person to attain feature extraction. More importantly, it can cope with slight variations, in terms of rotation, translation, and size difference, in images captured from the same person. The experimental results show that the performance of palmprint-based unimodal system and speech-based unimodal system fails to meet the requirement. Fusion at the matching-score level is used to improve the performance of the system. The psychological effects of such multimodal system should also not be disregarded and it is likely that a system using multiple modalities would seem harder to cheat to any potential impostors.

In the future we plan to test whether setting the user specific weights to different modalities can be used to improve a system's performance.

References

[1] A. A. Ross, K. Nandakumar, and A. K. Jain. Handbook of Multibiomtrics. Springer-Verlag, 2006.

[2] Mahesh P.K. and M.N. Shanmukhaswamy. Comprehensive Framework to Human Recognition Using Palmprint and Speech Signal. *In Springer-Verlag Berlin Heideberg 2011*.

[3] Mahesh P.K. and M.N. Shanmukhaswamy. Integration of multiple cues for human authentication system. *In Procedia Computer Science*, Volume 2, 2010, Pages 188-194.

[4] Jr., J. D., Hansen, J., and Proakis, J. Discrete Time Processing of Speech Signals, *second ed. IEEE Press, New York, 2000.*

[5] O. Rioul and M. Vetterli, "Wavelets and Signal Processing, "IEEE Signal Proc. Magazine, vol. 8(4), pp. 11-38, 1991.

[6] D. A. Reynolds and R. C. Rose, "Robust Text_Independent Speaker Identification Using Gaussian Mixture Speaker Models" IEEE Transactions on SAP, vol.3, pp, 72-83, 1995.

[7] P. Alexandre and P. Lockwood, "Root cepstral analysis: A unified view: Application to speech processing in car noise environments," Speech Communication, v.12, pp. 277-288,1993.

[8] D. A. Reynolds, "Experimental Evaluation of Features for Robust Speaker Identification," IEEE Transactions on SAP, vol. 2. Pp. 639-643,1994.

[9] Turk and A. Pentland, "Face Recognition using Eigenfaces", in Proceeding of International Conference on Pattern Recognition, pp. 591-1991.

[10] Turk and A. Pentland, "Face Recognition using Eigenfaces", Journals of Cognitive Neuroscience, March 1991.

[11] A. K. Jain, K. Nandakumar, & A. Ross, Score Normalization in multimodal biometric systems. *The Journal of Pattern Recognition Society, 38*(12), 2005, 2270-2285.

DOI: 01.IJSIP.03.01.7